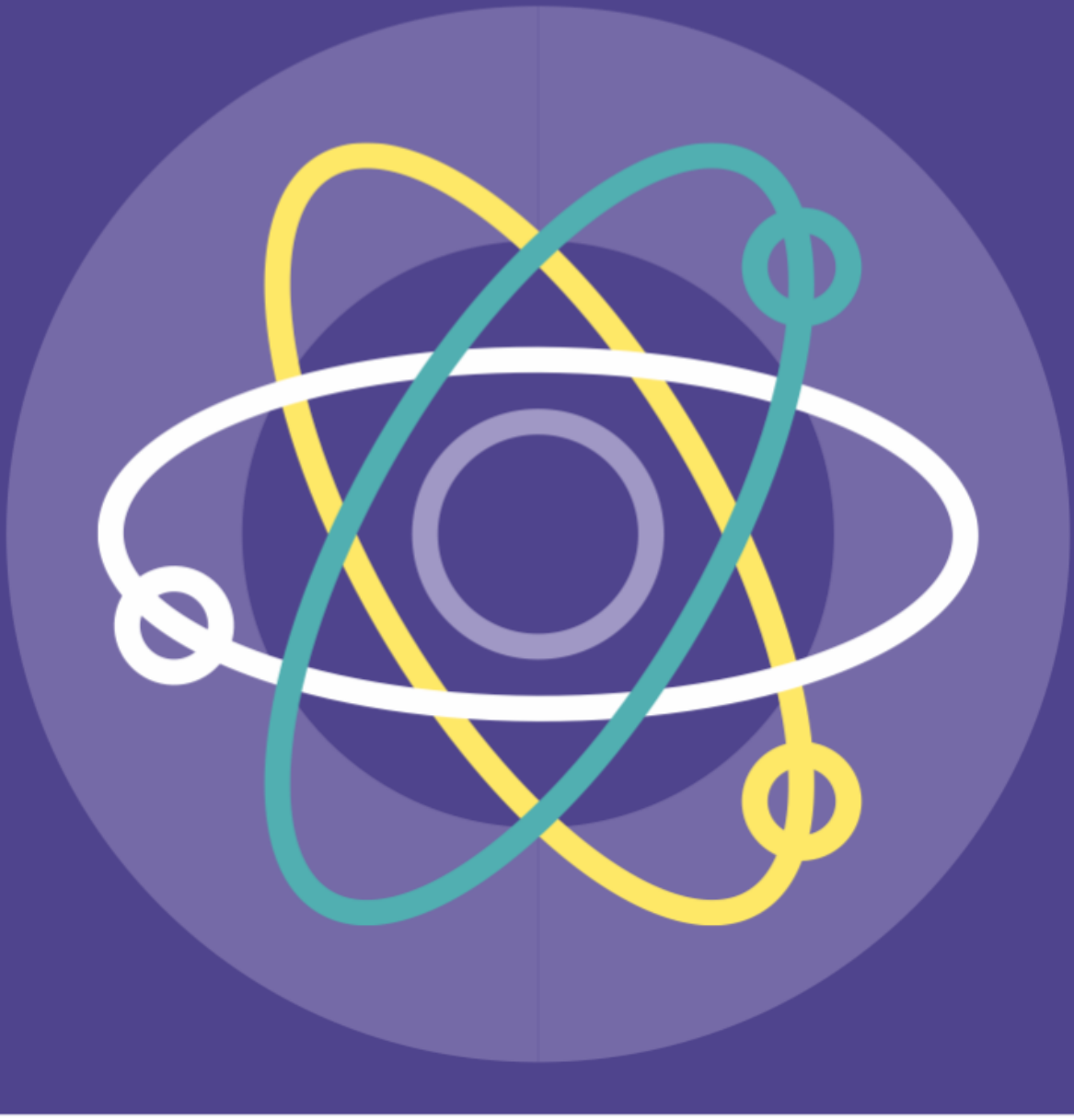
MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'ESPACE

Liberté
Égalité
Fraternité

Ouvrir la science !

Partager les données liées aux publications scientifiques

GUIDE POUR LES AUTEURS



Partager les données liées aux publications scientifiques

Guide pratique à l'usage des auteurs

———

Comité pour la science ouverte

Collège des Données de la recherche

———

Frédéric de LAMOTTE – Pilotage
INRAE

Véronique STOLL – Pilotage
Observatoire de Paris-PSL

Edition de mars 2022 :

Baptiste CECCONI
Observatoire de Paris-PSL

Jean-Yves CHATELIER
INERIS

Bénédicte KUNTZIGER
CCSD - CNRS

Yvette LAFOSSE
Inist-CNRS

Jean-François MARTIN
Institut Agro

Kenneth MAUSSANG
Université de Montpellier

Claire SOWINSKI
Inist-CNRS

Corentin SPRIET
CNRS

Coralie WYSOCZYNSKI

Carlo Maria ZWÖLF
Observatoire de Paris-PSL

Mise à jour de juin 2025 :

Christine HADROSSEK
DDOR

Hélène JOUGUET
Huma-Num

Emilie LERIGOLEUR
GEODE CNRS

———

Décembre 2025

———

Conception graphique : opixido

Sommaire

En matière de recherche scientifique, les publications constituent des vecteurs classiques de diffusion des connaissances. Elles reposent de plus en en plus sur des données et des analyses sous-jacentes. Le dépôt des données associées aux publications dans des entrepôts de données est aujourd'hui fortement encouragé voire imposé par les éditeurs de revues, les établissements, les agences de financement etc. Le partage de ces données est donc un élément important participant de la qualité des travaux de recherche. L'objectif de ce guide est de vous familiariser avec les étapes du partage des données liées aux publications en privilégiant une approche concrète. *Alors, on s'y met !?*

1 | Pourquoi partager les données liées aux publications ?

L'ouverture des données favorise la **transparence et la reproductibilité** du processus scientifique. En effet elle assure leur disponibilité pour tous et favorise leur réutilisation, permettant ainsi plus de transparence. Elle permet également un accès direct de ces produits de la recherche aux citoyens, contribuant ainsi aux débats publics consacrés aux enjeux sociétaux.

Ouvrir les données, c'est permettre leur réutilisation et donc les exposer à la critique. Pour s'y préparer, la stratégie consiste à les documenter mais aussi à mettre en place, tout au long de leur cycle de vie, des actions de gestion visant à préserver leur **qualité**. Les données ainsi mises à disposition bénéficient d'une **traçabilité** accrue et augmentent leur potentiel de réutilisation y compris pour leur producteur.

La diffusion ouverte des données assure la **reconnaissance** de leurs producteurs et leur **visibilité** ainsi que celle de leur établissement, au même titre que la publication assure la visibilité de ses auteurs. L'ouverture d'un jeu de données augmente sa visibilité et donc ses chances d'être utilisé par un autre projet de recherche puis cité de manière analogue à une publication. Les personnes qui y sont associées voient leur **implication valorisée**. Il est également à noter qu'une publication accompagnée de données est **davantage citée**[1].

Les données produites pendant un projet de recherche peuvent avoir de la valeur et un intérêt au-delà du projet, et parfois de la discipline initiale. Les rendre disponibles permet d'exploiter pleinement leur potentiel, favorisant ainsi l'**interdisciplinarité** et **la collaboration** académique.

2 | Comment partager des données liées aux publications ?

Lors de la rédaction d'un article scientifique, les auteurs adoptent naturellement une approche pédagogique consistant à bien définir toutes les notations et conventions utilisées dans l'article, à détailler les hypothèses, le cadre de travail ainsi que l'état de l'art afin d'en faciliter la lecture et en permettre la compréhension. La diffusion des données fait partie de cette approche. Si l'on veut que les données partagées puissent être utiles à la communauté, les mêmes attentions doivent être portées à la publication des données, y compris en amont de leur partage.

Préparation et documentation des données

[1] Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B (2020) The citation advantage of linking publications to research data. PLOS ONE 15(4): e0230416. https://doi.org/10.1371/journal.pone.0230416

Avant de procéder au dépôt des données dans un entrepôt, il est recommandé de préparer les informations et documents associés à leur production : dictionnaire des données expliquant les variables, protocoles expérimentaux, informations sur la provenance (genèse des données, traçabilité), hypothèses ou contraintes liées à leur production…. L'ensemble de ces informations constituent les métadonnées qui peuvent être génériques ou propres à certaines communautés[2]. Ces éléments favoriseront la compréhension et la réutilisation des données.

Parmi les bonnes pratiques, la gestion des données et les modalités de leur diffusion (choix de l'entrepôt) peuvent être identifiées et anticipées dans un plan de gestion de données[3]. Ce document évolutif permet de se poser les bonnes questions et d'anticiper les questions éthiques et juridiques en lien avec la politique d'ouverture des données de son l'établissement. Par exemple, dans le cas des données à caractère personnel, des règles de mise en conformité selon le Règlement Général pour le Protection des Données (RGPD) sont souvent nécessaires comme l'anonymisation ou la pseudonymisation des données et/ou des demandes d'autorisation de diffusion. Il est préconisé de se rapprocher du Délégué à la protection des données (DPD ou DPO) de l'établissement de la direction d'unité pour une inscription au registre des traitements.

Il est fortement recommandé d'utiliser des entrepôts de données institutionnels, généralistes ou disciplinaires pour le partage des données. Ces plateformes offrent un environnement spécialement conçu pour la documentation, l'ouverture et la réutilisation des données de recherche. Établir correctement le lien entre le jeu de données déposé dans l'entrepôt et l'article disponible sur une plateforme de publication devient alors une nécessité et une démarche à anticiper. De plus, en créant des liens entre la publication et les jeux de données ainsi partagés, vous améliorez le référencement de vos productions par les moteurs de recherche. Vos travaux seront donc plus facilement repérés, lus et cités.

Il est recommandé de ne pas confier les données à partager aux éditeurs des revues qui proposent de les déposer dans l'entrepôt de la revue elle-même ou sous la forme de *supplementary data* ou de *supplementary materials* associés à l'article. Une telle publication se fait encore trop souvent dans un format et un environnement qui ne permettent pas de documenter correctement les données et rendent difficile leur réutilisation. Elle peut aussi s'accompagner d'une demande de transfert exclusif de droits contraire à la loi française et à l'esprit de la science ouverte. Enfin, dans certains cas, elle contribue à rendre les utilisateurs captifs au sein d'environnements maîtrisés par de grands acteurs commerciaux de l'édition scientifique.

Des services d'appui à la recherche existent dans de nombreux établissements, ils offrent un accompagnement de proximité. Le réseau des ateliers de la donnée de l'écosystème Recherche Data Gouv constitue un point d'entrée bien identifié.

Choix de l'entrepôt

- Dans le cas de disciplines structurées pour le partage des données (astronomie, génomique, etc…), les producteurs de données ont à disposition des entrepôts spécifiques à leur discipline ;
- Ces entrepôts sont répertoriés dans des catalogues comme Cat OPIDoR[4] et re3data[5]. Pour guider les équipes de recherche, le Collège des données de recherche du Comité pour la science ouverte propose une méthodologie pour identifier les entrepôts de confiance, ainsi qu'une liste descriptive de ces entrepôts permettant l'auto-dépôt[6][7].
- A défaut d'entrepôt disciplinaire, les producteurs de données pourront se tourner vers l'entrepôt pluridisciplinaire Recherche Data Gouv et ses espaces institutionnels.

[2] Métadonnées, standards, formats : fiche synthétique. Doranum, https://doi.org/10.13143/vbjs-6288

[3] La minute plan de gestion de données. Doranum, http://doi.org/10.13143/dwmf-2j16

[4] Cat Opidor, wiki des services dédiés aux données de la recherche : https://cat.opidor.fr/

[5] Re3data (REgistry of REsearch Data REpositories) : https://www.re3data.org/

[6] Frédéric de Lamotte, Véronique Stoll, Cécile Arènes, Marie-Emilia Herbet et al.. Sélectionner un entrepôt thématique de confiance pour le dépôt de données : méthodologie et analyse de l'offre existante. Comité pour la Science Ouverte. 2024. DOI 10.52949/52 .

[7] Liste des entrepôts thématiques de confiance permettant l'auto-dépôt accessible via https://www.ouvrirlascience.fr/selectionner-un-entrepot-thematique-de-confiance-pour-le-depot-de-donnees-methodologie-et-analyse-de-loffre-existante/

Il est recommandé d'**opter pour un entrepôt intégrant des métadonnées structurées**, et assurant la curation des dépôts, afin de garantir l'interopérabilité, la qualité descriptive et la réutilisabilité des données déposées.

La plateforme nationale des données Recherche Data Gouv

Inaugurée en 2022, Recherche Data Gouv est une infrastructure nationale dédiée à la gestion, au partage et à l'ouverture des données de la recherche française. Elle s'appuie sur une plateforme centrale pour la préservation, la publication et la découverte des jeux de données, et sur un réseau d'accompagnement de proximité pour soutenir les équipes de recherche dans la gestion et la diffusion des données. Recherche Data Gouv garantit la souveraineté, l'accessibilité et la réutilisabilité des données grâce à des licences adaptées et des métadonnées normalisées.

Le choix d'un entrepôt de confiance garantit notamment :

- L'assignation d'un identifiant persistant (*Persistent IDentifier* - PID, ex : DOI, ARK...) pour la citation de vos données, ce qui constitue le socle de base pour établir le lien avec d'autres produits de la recherche comme les publications ;
- La description des données (métadonnées) à un niveau suffisant pour en faciliter la découverte, la compréhension et la réutilisation ;
- L'utilisation de licences et la définition de règles d'accès permettant d'inscrire la réutilisation dans un cadre légal bien défini et compatible avec le droit français et européen[8] ;
- Une durée de conservation minimale de plusieurs années, cohérente avec la politique de conservation des données de l'établissement de rattachement.

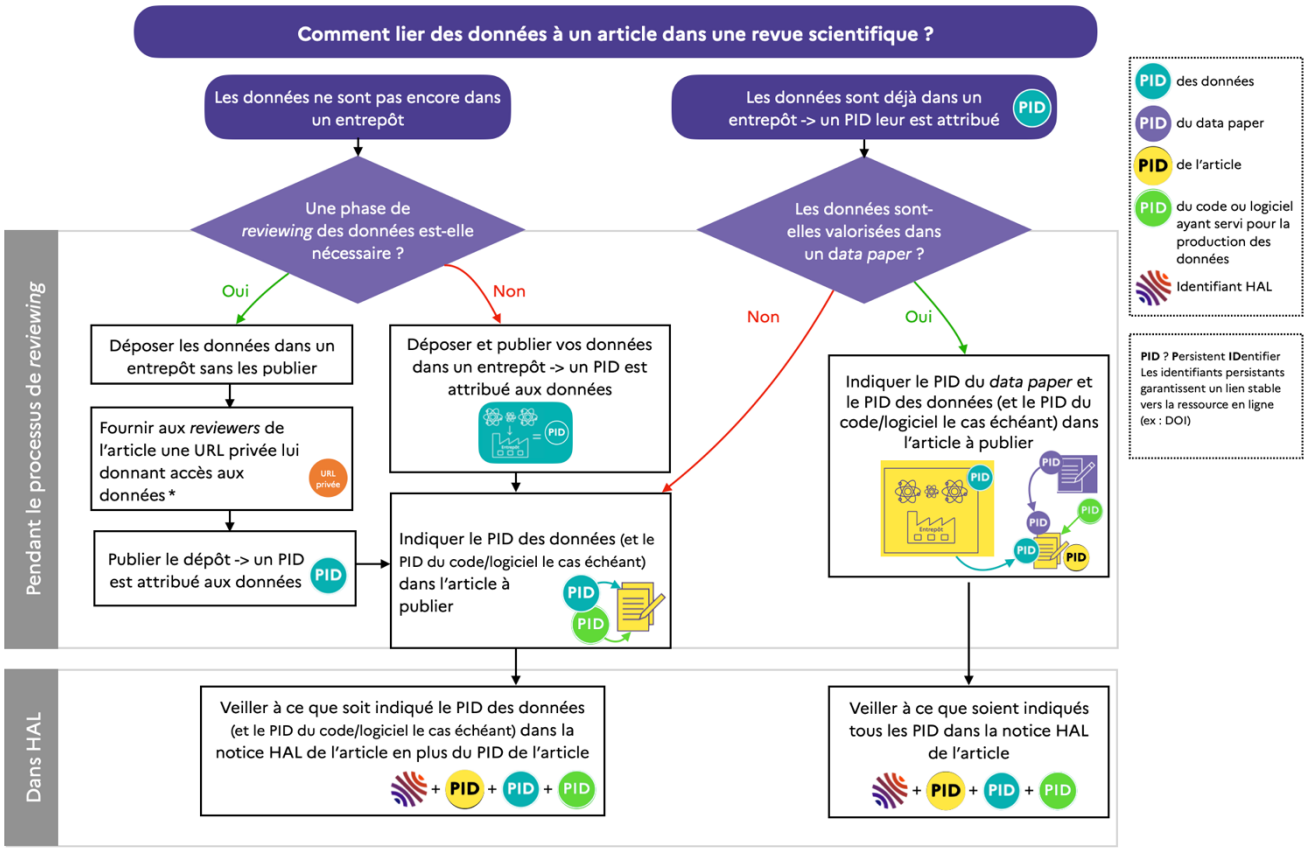
Lier les données aux publications

Plusieurs options sont disponibles pour établir la liaison entre un article et des données qui lui sont associées **avant ou pendant le processus de publication de l'article** considéré. Il est alors facile de créer le lien entre l'article et les données associées, selon les modalités décrites dans le schéma à la page suivante. De même, le référencement des publications associées aux données (y compris les *data papers*) est généralement possible dans tous les entrepôts de données, même après le dépôt initial.

À l'inverse, indiquer le lien vers les données dans un article **après sa publication** est souvent difficile, voire impossible, à l'heure actuelle. Une solution consiste à faire référence aux identifiants persistants des données dans la version de l'article déposée dans une archive ouverte (comme HAL par exemple). Cela permet ainsi la liaison réciproque entre publication et données, mais seulement pour la version déposée dans l'archive ouverte. À noter que la tendance est à l'automatisation du référencement réciproque de ces PID dans HAL et dans l'entrepôt hébergeant les données associées.

[8] Les licences de réutilisation dans le cadre de l'Open data et de la loi pour une République numérique : https://doi.org/10.13143/ssh2-zd93

Comment lier des données à une publication lors du processus de *reviewing* de l'article (révision par les évaluateurs de la revue) et comment les référencer dans la notice HAL de l'article.



*Dans le cas où l'entrepôt ne fournit pas d'URL privée, vous pouvez communiquer le PID des données publiées dans l'entrepôt ou transmettre les données au *reviewer* par un autre moyen

Les *data papers*

Un *data paper* est une publication dont le but est la description d'un jeu de données de la recherche. Contrairement à un article de recherche classique, le *data paper* consiste en une description détaillée des données scientifiques, des métadonnées, ainsi que les circonstances et méthodes de leur collecte, mais sans analyse ou interprétation de ces données. L'objectif est de permettre à toute personne intéressée d'exploiter ces données. Le jeu de données décrit doit être accessible, déposé dans un entrepôt approprié, et muni d'un identifiant persistant de type DOI.

Un *data paper* est publié sous forme d'un article revu par les pairs, gage de sa qualité, et peut être cité au même titre qu'un article classique. Par conséquent, l'auteur d'un *data paper* doit être convaincant quant à la qualité et à la portée scientifique des données (notamment leur potentiel de réutilisation). Le *data paper* peut être publié dans des revues spécifiques (data journals) ou dans des revues scientifiques traditionnelles qui permettent ce format[9].

Les codes et logiciels

Les codes et logiciels sont des productions de la recherche concernés par les principes de la science ouverte. Consultez le livret « Science ouverte : Codes et logiciels » pour répondre à toutes vos interrogations ! [10]

Lier des ressources (jeu de données, publications, codes sources, images, sons…) à une publication dans HAL

L'archive ouverte de documents scientifiques HAL permet de signaler facilement les liens entre des publications et d'autres ressources, comme un jeu de données, un code source archivé sur Software Heritage ou un autre dépôt HAL[11]. La section du formulaire de dépôt nommée "Ressources associées" permet d'ajouter une ou des relations entre un dépôt dans HAL et un jeu de données ou un code source à la condition que la ressource associée soit dotée d'un identifiant persistant (DOI, SWHID…). A noter que le partenariat avec Recherche Data Gouv et Software Heritage permet, de plus, de visualiser dans la notice HAL, pour les données issues de ces entrepôts, la donnée et de ses métadonnées. Cela favorise le partage et la réutilisation des données par la communauté scientifique, augmentant ainsi l'impact de la publication[12].

Citer un jeu de données

La façon de citer un jeu de données lié à une publication scientifique dépend des circonstances de production de ces données :

- Si les données ont été produites et partagées à l'occasion de la rédaction de l'article, il est recommandé d'introduire une section spécifique « Disponibilité des données » avant les références bibliographiques. Par exemple :

  Disponibilité des données : Les jeux de données liés à cet article peuvent être trouvés sur https://doi.org/10.23708/PQTQDA, hébergé par DataSuds IRD (Granjon and Fossati, 2020)

[9] Comment trouver des-data journals ? : https://doi.org/10.13143/2q45-cg36

[10] https://www.ouvrirlascience.fr/science-ouverte-codes-et-logiciels/

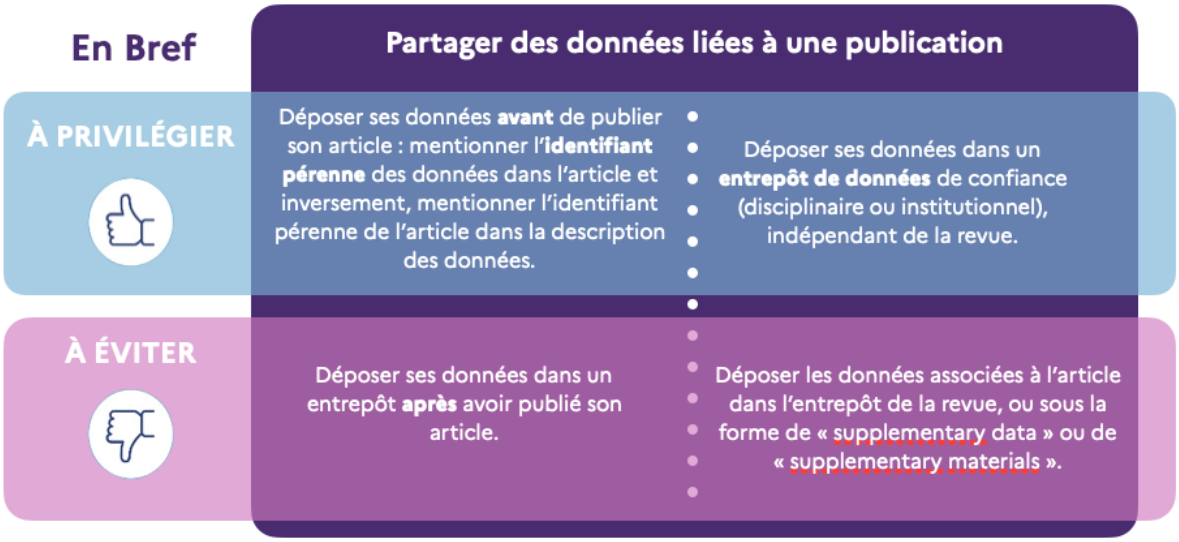
[11] https://hal.science/

[12] https://doc.hal.science/relations-depots/

- Si les données ont déjà été produites et partagées dans un autre cadre que celui de la publication, la citation se fait dans les références sous une forme équivalente à celle des références bibliographiques, par exemple :

  Van Halder, Inge ; Sacristan, Alberto ; Martín-García, Jorge ; Pajares, Juan Alberto ; Jactel, Hervé, 2022, « Monochamus g alloprovoncialis catches and pine tree composition in different landscape buffers in Spain », https://doi.org/10.15454/JXFGPI, Portail Data INRAE, V1

La citation correcte des données permet une meilleure indexation et donc une meilleure découverte lors de la recherche et donne un crédit permanent au producteur des données.



Pour aller plus loin :
Le guide « Science ouverte – Données de la recherche » 2024 aborde les principales notions relatives à la gestion et à la diffusion des données de la recherche : https://www.ouvrirlascience.fr/science-ouverte-donnees-de-la-recherche/

Glossaire

Données de recherche : Informations collectées, observées, générées ou créées dans le cadre d'un processus scientifique, et qui servent de base à l'analyse, à la validation ou à la reproduction des résultats de recherche. Elles peuvent prendre des formes très diverses selon les disciplines : fichiers texte, images, sons, vidéos, descripteurs d'objets physiques, bases de données… Les données de recherche peuvent être brutes, traitées ou analysées, et incluent également les outils et protocoles utilisés pour les décrire.

Entrepôts de données : Infrastructure de stockage et de services permettant le dépôt, la description, le partage en accès ouvert, la découverte et la réutilisation, par des humains ou des machines, de jeux de données. Ces jeux de données sont associés à des métadonnées et sont conservés à moyen ou long terme.

FAIR : ensemble de principes visant à soutenir la recherche en facilitant la réutilisation des données. Faciles à trouver (*Findable*), Accessibles (*Accessible*), Interopérables (*Interoperable*), Réutilisables (*Reusable*). Pour aller plus loin : https://www.ouvrirlascience.fr/fair-principles/

Indexation : attribution à un document de termes distinctifs (des mots-clés par exemple) renseignant sur son contenu et permettant de le retrouver.

Licence : texte juridique définissant les conditions de diffusion et de réutilisation d'une production (par exemple : licences logiciels libres, *Creative Commons*).

Métadonnées : Ensemble d'informations structurées qui décrit, explicite, localise une ressource informationnelle, dans le but d'en faciliter la recherche, l'usage, et la gestion. Pour aller plus loin : https://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf

Open data : données ouvertes, dont l'accès est libre et sans restriction. Pour aller plus loin : https://sparceurope.org/new-sparc-europe-report-analyses-open-data-open-science-policies-europe/

PID (Identifiant persistant ou Persistent IDentifier (PID) : Chaîne de caractères alphanumériques qui identifie de façon non ambiguë et persistante dans le temps un objet numérique ou physique (ex: DOI Crossref pour une publication, DOI Datacite pour une donnée, SWHID pour un code ou logiciel, ORCID pour un contributeur, ROR pour une structure...).